NeSA DRC Overview for TAC – February 17, 2009

Field-testing Plan

Proposed Design for NeSA: Develop three operational forms

Three operational forms (Form A, Form B, and Form C) will be developed per content area: mathematics, reading, and science**. Note: The initial standalone reading field test in spring 2009 will not yield enough items to construct three forms. As a result, the proposal includes embedded field testing for reading in the first operational year (2010) in order to have enough items to construct the third form. It is estimated that mathematics and science standalone field testing will yield enough items for the construction of three operational forms. For reading, mathematics, and science, the three forms will be parallel in terms of the content test blueprint. There may be some overlap of items across operational forms, if needed.**

Reading

Reading will be operational in 2010, with the FT in 2009 (Two of the three unique forms of the operational tests will be constructed from the FT in 2009). One (Form A) of the two operational forms will be administered in 2010. There will be six forms of Form A administered in 2010, each with a unique core set of items. Each form will also have a set of 10 embedded field test items. The embedded field testing in 2010, along with some items left in the bank from the initial field test in spring 2009, will provide for enough items to construct the third form, Form C.

Mathematics

Mathematics will be operational in 2011, with the FT in 2010. (We need three forms of the operational tests for 2011 and all three forms will be constructed from the original standalone field test.)

Science

Science will operational in 2012, with the FT in 2011. (We need three forms of the operational tests for 2012 and all three forms will be constructed from the original field test.)

It is important to note that the third form would not be printed and/or administered under the contract, but would be constructed.

Scaling and Equating Plan

Item Calibration and Linking

Item calibration and linking will be accomplished via a common calibration of all forms (Winsteps, 2008). This in effect treats the entire field test as a single form with many item responses missing. This is a very efficient method to create a common logit metric across multiple forms. The linking can be cross validated by assuming randomly equivalent samples of students and using *common person equating*.

With Rasch calibration, obtaining a sample representative of a larger population is not a major issue. However, it is prudent to use a broad selection of students and important to ensure the inclusion of all subpopulations. Both the computer-based and the paper-based assessments will depend on samples that are voluntary. The method by which the paper-based sample was created ensures that group matches the state demographics as well as practicable.

Because the computer-based group is much larger, it can be sampled to create calibration groups that are representative. If this process is replicated multiple times, it can also be used to generate estimates of standard errors.

The common item calibration will result is a calibrated pool of items. Any selection of items from the pool will provide an equated test form. Equating in this sense means that scale scores derived from any such form will relate to same measurement scale and can be compared to any scale scores from the pool or to performance levels developed for the pool.

The field test and link design is expected to provide a calibrated pool adequate to construct the two initial operational forms with six passages with approximately seven items each. The calibrations will allow the forms to be equated prior to the operational administration and the raw-to-scale score conversion tables prepared immediately after the operational forms are finalized and once the scale score metric has been defined.

Scale Scores

Establishing the final Scale Score metric is typically done after setting the performance standards. Scale Scores are a linear transformation of the logit metric, i.e., $Scale\ Score = A + B*Logit$, where A and B are chosen to make the metric easier to interpret. Scale scores retain all the measurement properties of logits although they are not used directly in the logistic functions.

The parameters A and B are generally chosen so that the metric:

- Does not involve negative numbers,
- Does not require decimal points to report at the student level,
- Fixes two points on the scale, or one point and the unit, which makes the scale easier to apply.

It may be desirable to choose a metric that does not resemble more familiar scales. Choosing $A = 50$ and $B = 10$, for example, would meet the first two conditions but probably not the third. The resulting metric would look like percent correct but would be interpreted very differently. In particular, item difficulty would be reversed with high numbers representing easy items and low numbers, difficult items. The table shows a few obvious choices and the scales with which they might be confused.

Table 3: Scale Score Definitions

A	B	Likely Range	Possible Confusion
50	10	0-100	Percent Correct
100	20	0-200	IQ
800	100	0-1600	SAT Total
500	91	0-1000	SAT
1000	144	200-1800	SAT Total

The last two entries under B are derived[1] from the natural logs of 3 and 2. Scaling by 91 will provide a scale such that a difference of 100 and 200 scale score points represent odds of 3:1 and 9:1 and, hence, probabilities of success of 0.75 and 0.90 respectively. Scaling by 144 would give probabilities of 0.67, 0.80, and 0.89 for differences of 100, 200, and 300.

A more popular strategy is to fix two meaningful points, perhaps the cut score for the *Proficient* and the cut score for the *Advanced* performance levels, at convenient round numbers. If the logit cut scores, as determined by some standard setting process, are c_p and c_a, and the scale score cut points that we would like are S_p and S_a, then A and B can be computed with:

- $B = (S_p - S_a) / (c_a - c_p)$, and
- $A = S_p - Bc_p$ or $A = S_a - Bc_a$.

One might also set the proficient cut score at S_p and the distance from *Basic* to *Advanced* at D. Then the calculations are:

- $B = D / (c_a - c_b)$, and
- $A = S_p - Bc_p$.

The actual numbers that are chosen to represent the Scale Scores have no inherent meaning *per se*. Attaching meaning to a given value for a score requires experience, just as knowing what jacket to wear when the local weather forecast predicts a high temperature of 25°C. Choosing A and B so that Scale Scores that correspond to meaningful events are convenient and easily remembered will help users acquire that experience.

NeSA Standard Setting/Standards Validation Plan

For the standard settings and standards validations required, DRC has 2 options for NDE's consideration:

1) DRC's original modified Bookmark standard setting and contrasting groups validation study as discussed in our original proposal, or

2) Contrasting Groups Study, with panel committee review.

[1] Specifically, ln (3) = 1 / 0.9102 and ln (2) = 1 / 1.4427.

Original Proposal Response

Bookmark

DRC proposed a plan to complete both the standard settings and standards validation of all subjects included in the NeSA assessments. Student level results would be reported indicating an overall level of performance according to the three achievement levels established by NDE. All standard settings will be conducted the summer following the first operational administration of each of the NeSA assessments. The table below, Table 4–11, details the schedule for these events.

Table 4–11. Standard Setting Schedule

Subject	Standard Setting Year	Standards Validation Years
Reading	2010	2010 and 2011
Mathematics	2011	2011 and 2012
Science	2012	2012 and 2013

DRC believes the standard settings for reading and mathematics should take a week and science should last three days. DRC plans to run the standard setting with three grade groupings (elementary, middle, and high school) concurrently starting with grades 5, 6, and 11 and ending with grades 3 and 8. Grade 11 in reading and mathematics would be completed in a shorter time frame, ending after three days. It is important that all panelists be trained together to maintain consistency and coherence.

DRC would use the Bookmark standard setting method to set standards for all subjects of the NeSA. The Bookmark procedure (Lewis, Mitzel, & Green, 1996) is appropriate for this project, as items can be reliably ordered by difficulty.

The proposed plan will use three separate panels: elementary (grades 3–5), middle school (grades 6–8), and high school (grade 11). Using the same panel for three consecutive grades will help ensure coherent recommendations.

When testing in consecutive grades, it is crucial that the performance standards be coherent across grades. Although this was not initially included in the development of the Bookmarking procedure, it is an important consideration in the procedure DRC is proposing. The process begins with grades 5 and 6. When these standards have been tentatively established by separate panels, DRC is proposing to bring the two panels together to discuss the work jointly. The final recommendations will then be developed with the input from the other panel. As the recommendations are developed for the remaining grades (grades 3 and 4 for the elementary panel and grades 7 and 8 for the middle school panel), the panels will be reminded of the joint results of the five-six panels to maintain a consistent pattern across grades.

Prior to the standard setting meeting, DRC proposes to involve all Nebraska teachers in a *contrasting groups* study. There are two purposes for this study. First, it will provide first-hand information from the classroom teacher about each student's expected performance on the assessment. Second, it provides a cost-effective strategy to validate the performance standards in

the second operational year for each content area. This will avoid the necessity, and all the associated costs and delays, of reconvening standard setting panels in the second operational year. The data collected in the first operational year will provide a basis for evaluating the second year results and to provide feedback and guidance to the panels during the item mapping process.

Alternative Method- Contrasting Groups with Committee

DRC would invite all Nebraska teachers, who teach the subject being assessed, to participate in a contrasting groups event the first operational year. As discussed above, this allows for first-hand information from the classroom teacher about each student's expected performance on the assessment. The survey will be made available to the teachers the same weeks of the student administrations.

Subject	Contrasting Group Study
Reading	2010
Mathematics	2011
Science	2012

In June, DRC would then convene a committee of the major stakeholders identified by the NDE, i.e. teachers, administrators, Special Education and English language specialists, legislators or business partners, who would meet for two days to discuss the results of the contrasting groups study and to set standards. There would be ten to twelve panelists that could be divided into upper and lower grades. This committee would:

- Review the assessments by grade
- Review contrasting groups data
- Review impact data
- Review PLD's

Reports

Timeline

DRC has proposed the following schedule for implementation of reporting by year:

Table 4–10. Proposed Report Deliverables and Timeline

	Online Parent Letters & District Student Data Files	Online Summary Reports	Hardcopy Parent/Guardian Reports
Year 1 (2009)	n/a	n/a	n/a
Year 2 (2010)	n/a	August 2010	August 2010
Year 3 (2011)	May 2011 (reading only)	August 2011	August 2011
Year 4 (2012)	May 2012 (reading & math only)	August 2012	August 2012
Year 5 (2013)	n/a	May 2013	May 2013

Given the field testing schedule and standard settings that occur for each operation subject, a two-week turnaround will not be accomplished until the 2013 school year. To get the maximum level of detail in the hands of parents and district personnel, DRC has proposed district student data files and online parent letters for years 2-4. District student data files will contain results for subjects for which standards have been set. They will contain data, organized by school, regarding student demographics and performance. The online parent letters will also contain scores for subjects where standards have been set. They can be printed and distributed by schools and districts to students and their families as an advanced opportunity for seeing results for that spring's assessment.

NeSA Reports

Electronic:	Classroom Roster
	Classroom Summary
	School, District, and State Report Packages
Hardcopy:	Parent/Guardian Report

Administration Training

Test administration training workshops will be presented by NDE and DRC/CAL March 23-25, 2009, for the spring 2009 reading field test administration. These sessions will address both the online and paper/pencil mode of testing.

In addition, five Web Ex training sessions will be conducted with district technology contacts March 4-12, 2009, with the focus being preparation for the online test.